D32–D36 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj014

ISfinder: the reference centre for bacterial insertion sequences

P. Siguier, J. Perochon, L. Lestrade, J. Mahillon[1] and M. Chandler*

Laboratoire de Microbiologie et Génétique Moléculaires, C.N.R.S., 118 Route de Narbonne, F-31062 Toulouse Cedex, France and [1]Laboratoire de Microbiologie Alimentaire et Environnementale, Université catholique de Louvain, Croix du Sud, 2 Bte 12, B-1348 Louvain-la-Neuve, Belgium

Received August 11, 2005; Revised and Accepted September 14, 2005

ABSTRACT

ISfinder (www-is.biotoul.fr) is a dedicated database for bacterial insertion sequences (ISs). It has superseded the Stanford reference center. One of its functions is to assign IS names and to provide a focal point for a coherent nomenclature. It is also the repository for ISs. Each new IS is indexed together with information such as its DNA sequence and open reading frames or potential coding sequences, the sequence of the ends of the element and target sites, its origin and distribution together with a bibliography where available. Another objective is to continuously monitor ISs to provide updated comprehensive groupings or families and to provide some insight into their phylogenies. The site also contains extensive background information on ISs and transposons in general. Online tools are gradually being added. At present an online Blast facility against the entire bank is available. But additional features will include alignment capability, PsiBLAST and HMM profiles. ISfinder also includes a section on bacterial genomes and is involved in annotating the IS content of these genomes. Finally, this database is currently recommended by several microbiology journals for registration of new IS elements before their publication.

INTRODUCTION

The massive accumulation of sequenced bacterial genomes over the past decade is providing exciting opportunities for understanding genome organization and evolution. This area of investigation depends on genome comparisons and therefore on accurate annotation of the genomes to be compared. In addition to open reading frames (ORFs) or potential coding sequences (CDS) and the signals for gene expression, correct annotation of other features such as mobile genetic elements (MGEs) is also essential. The bacterial genome is composed of a core minimal species genomic backbone decorated with a variety of additional elements. These MGEs include bacteriophages, conjugative transposons, integrons, unit transposons, composite transposons and insertion sequences (ISs) (1). They can be found in different combinations in different bacterial species or in different isolates of one species and form part of an extensive horizontal gene pool. Many bacterial genomes can be considered as a mosaic of these different elements.

MGEs often represent a significant proportion of the host genome and are intimately involved in directing gene exchange and reassortment. Unfortunately, in many cases, annotations include only the ORFs or potential coding sequences carried by the elements and ignore their DNA ends, which are an essential feature of their activity in mediating gene rearrangements. Moreover, the presence of IS vestiges is rarely annotated and the corresponding scars of ancestral rearrangements are thus often obscured. Similarly, mobile cassettes missing the transposase gene (2,3) can be easily overlooked. The availability of specialized dedicated databases is a powerful tool in this type of analysis.

ISfinder is a database dedicated to bacterial ISs which constitute one of the largest groups of MGEs (4). These are relatively short genetically compact DNA segments (between 0.7 and 2.5 kb) encoding no functions other than those involved in their mobility. Many, but not all, carry short (<40 bp) imperfect inverted repeats (IR) at their ends and generate a small (between 2 and 14 bp) duplication of the target DNA flanking the point of insertion (DR). At present, an estimated 1500 different ISs have been detected. They have been observed in most bacterial genomes and plasmids where they may be present in high numbers.

THE SITE

ISfinder (www-is.biotoul.fr), originally conceived as a tool for classifying IS elements, has now also assumed the role

*To whom correspondence should be addressed. Tel: +33 5 61 33 58 58; Fax: +33 5 61 33 58 86; Email: mike@ibcg.biotoul.fr

of a reference centre (Fig. 1) following the closure of the University of Stanford centre. It is limited to IS elements and does not yet include other transposable elements. It is divided into several sections.

Finding general information

Extensive background information on ISs and transposons in general is included in *Information*/*General information*. This is based on two broad reviews (4,5) which have been enhanced with supplementary material including a large bibliography and a series of colour figures. The feasibility of including animated images of different transposition mechanisms for teaching purposes is being studied. Information concerning their regulation, structure, catalytic mechanism and target site specificity is presented in subsections.

Finding information on specific IS families

ISs can be grouped into relatively distinct families based on their genetic organization, the similarities between their transposases and the relationship of their IRs (4). The characteristics of each family are also described in *Information*/*IS families*/*Major features of prokaryote IS families* as is the nature of the catalytic properties of their transposases (/*The DDE motif*) and other more detailed individual family characteristics (/*Family information*) which includes a general section (*Occurrence, Variety and Systematics*) and sections on individual families. There is also a large bibliography (/*References*).

This grouping is an evolving aid to classification and to managing the high number and variety of ISs, which are being identified in the various genome sequencing projects. There are links in the text to descriptions of each family. The section is updated periodically.

IS nomenclature

Another role of ISfinder is to assign IS names and to provide a focal point for a coherent nomenclature. The section *Information*/*Nomenclature and Attribution requests*/ includes an explanation of the nomenclature scheme and suggestions for nomenclature for different bacterial species (/*IS Nomenclature*). We have adopted a nomenclature similar to that of restriction enzymes. Although this is not perfect, since some ISs are found in different species or even in different genera, the system is viable and has the advantage of indicating the host species rather than being confronted with a long series of numbers in the names as in the original nomenclature system (6).

Originally, blocks of IS numbers were assigned to individual scientists, groups or institutions (6). The site includes a listing of these original ISs numbers together with the last known address of the attribution [/*Reserved blocks of IS numbers previously attributed (Stanford University listing)*].

A third subsection lists the names of ISs which have been attributed by ISfinder or which have appeared in the literature in a form which did not correspond to the recommended nomenclature (/*List of IS names currently attributed*). The attribution address is also included.

Getting an IS name

ISfinder includes an online form for registering (requesting a name for) new ISs (/*Attribution requests*). Several journals now suggest that authors register their new IS elements with ISfinder before publication. Since a unique name can be attributed, this avoids some of the confusion in the literature. The form includes the following fields: BACTERIAL HOST, (e.g. *Streptococcus pneumoniae*); REGISTRANT; Your email; Title; Name; First name; Institution; Address; City; State and Zip code; Country; Comments. The registrant receives an email with the attributed IS name. At present the reply is manual but shortly will be automatic.

THE DATABASE

Using the database/

ISfinder is also the repository for ISs. Each new IS is indexed together with information such as its DNA sequence and ORFs or potential coding sequences, the sequence of the ends of the element and target sites, its origin and distribution and family attribution, together with a bibliography where available. Information on each IS is stored as an individual file in MySQL format with links to the NCBI Taxonomy Browser and to the relevant public database file through its accession number.

Finding an IS

ISfinder can be searched using the online search tool (*Using the database*/*search*/). This is available as a simple or a more advanced search. In the simple search, the */Output Layout* can be defined to extract different types of information (Files: an entire file with all information concerning the IS of interest; origin: the original bacterial host; listing: a table of all ISs found in the search together with various characteristics such as Synonyms, Isoforms, Family, Origin, Accession Number, Length, IR and DR; hosts: other bacterial species which also contain the IS; references: a short bibliography on the IS; comments: general comments). The type of result obtained through */Output Layout* depends on the criteria requested in */Search in all fields* (*Name; synonyms; Iso; Family; Origin; Comments; References*). A filter has been added to allow more flexible searching (contains, equal to, begin with and end with).

Getting more extensive information

The *extensive search* allows a larger set of *Output layouts* (Files; GCG; origin; listing; families; ORF; insertion sites, IR; IRL-IRR; iso; hosts; references; and comments) together with an extended */Search in all fields* [*Name; Synonyms; Iso; Family; Group; Origin; Hosts; Accession Number; Length (in bp); Insertion Sites; IR; DR; ORF; Left End; Right End; Comments; References; IS-SEQ; IS-PEP*].

Online tools are gradually being added (*Using the database*/*analysis*/). At present an online Blast facility (/*BLAST*) against the entire bank is available. Future features will include alignment capability, PsiBLAST and HMM profiles.

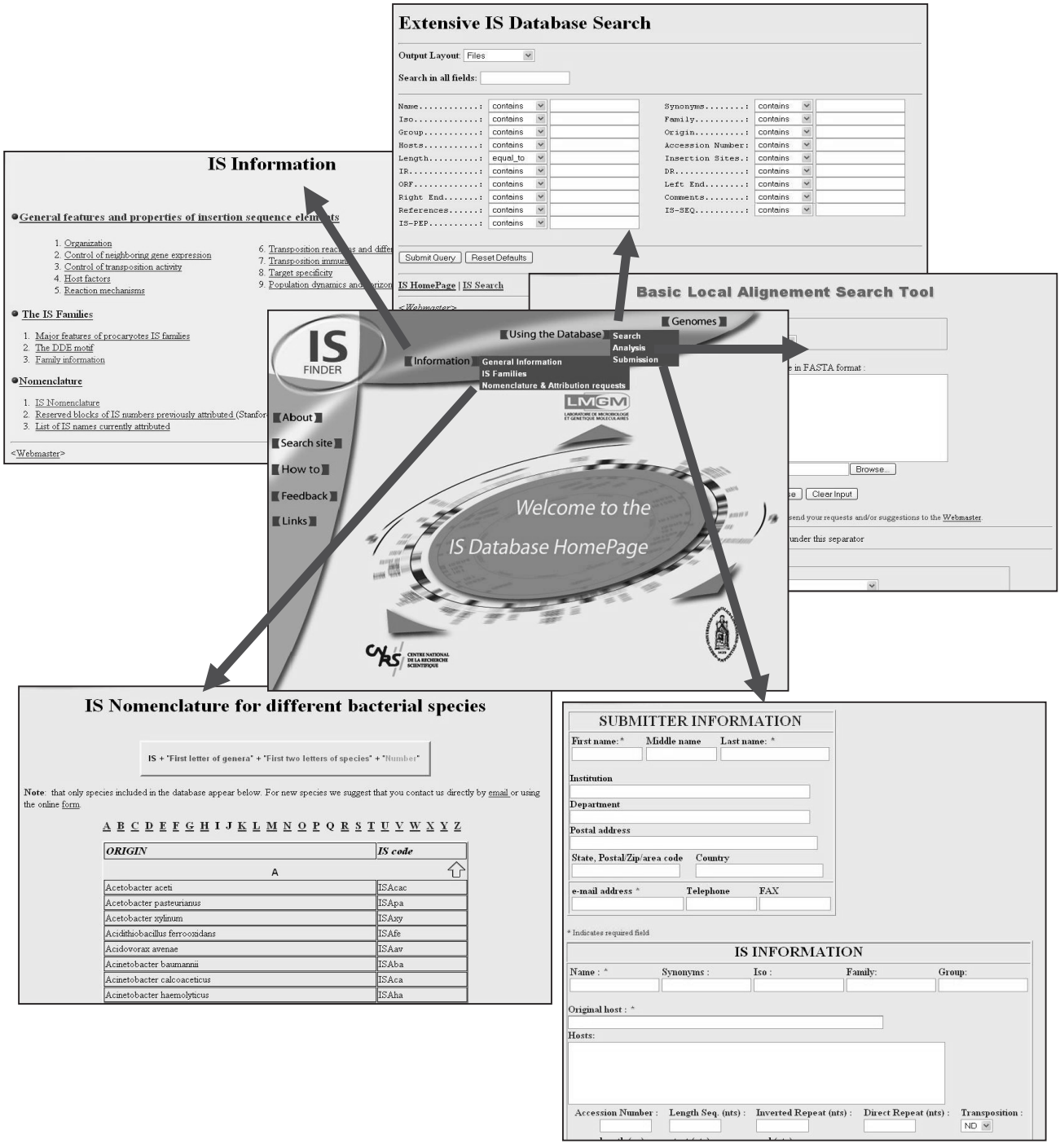

Figure 1. An overview of ISfinder showing links from the home page to various pages.

Submitting your IS to the database

Following the attribution of a name, we ask that the sequence be deposited in the database (*Using the database*/*submission*/). This enriches the base and helps subsequent users. It also enables us to maintain an overview to prevent multiple names being attributed to a single IS. An online form is available for this and minimally involves completing the boxes marked '*'. This includes both institutional information to post in ISfinder and important information concerning the IS.

Special simplified submission procedures can be arranged for deposition of whole genome complements of ISs. Following online submission, each submitted sequence file is verified before inclusion into the public database. Confidential sequence information is retained in a secure database prior to being released (following accord) into the public domain. An automatic reply will be instituted informing the person submitting the IS that it has been added to the public database.

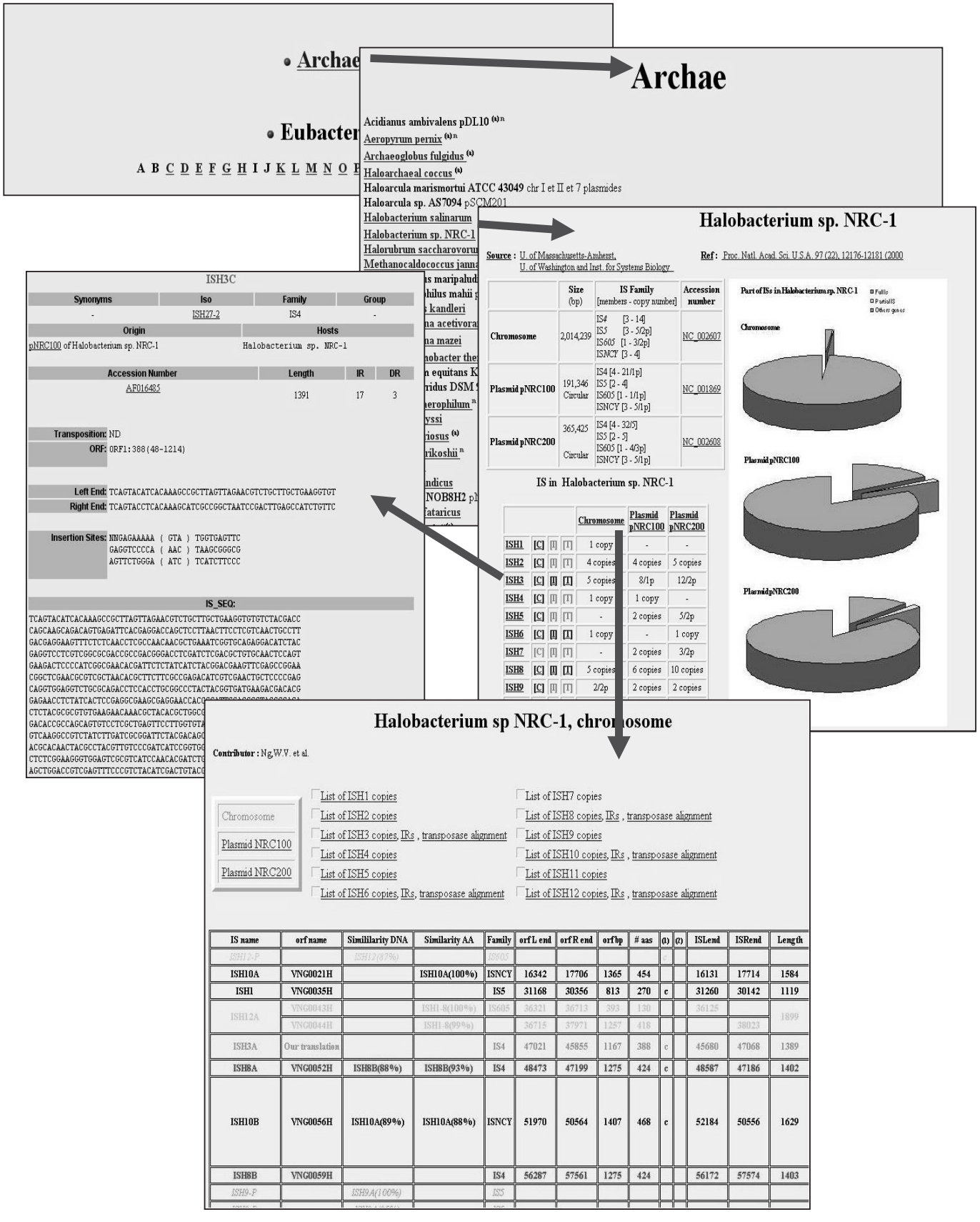

Figure 2. The Genome pages. This figure presents the Archea *Halobacterium* sp. NRC-1 as an example.

Genomes

ISfinder includes a section on bacterial genomes. In view of the frequency of newly published genome sequences, this section is not complete. It includes sections on both chromosomes and plasmids. Genomes for which extensive analysis has been undertaken are linked to the appropriate information.

The genome section is at present being restructured. Eubacterial and archeal genomes are separated into two sections from which a list of available genome sequences is accessible.

Each bacterial name gives access to a page containing information (name, length, ISs, Accession number, source and references) concerning ISs from all sequences genomes of a given species (Fig. 2). Each IS is represented by an individual file in the database together with information concerning its location(s) together with partial copies, a list of the different copies with their DNA sequences, alignments or their terminal IRs where relevant and alignments of their putative transposases.

CONCLUDING REMARKS

ISfinder has been operational for several years and we expect an increasing number of online submissions both from individuals (an aspect which at present functions relatively well) and especially from the genome sequencing projects (which at present involves only a limited number of sequencing centers).

One general goal of ISfinder will be to interact with other complementary specialized databases such as those including bacteriophages, plasmids, integrons, recombinases and genomic islands. One ongoing project is to provide an interface with ACLAME (A CLAssification of genetic Mobile Elements: http://aclame.ulb.ac.be/) (7).

Finally, ISfinder functions as a research tool. Thorough and systematic bacterial genome analysis has already identified several phylogenetically related groups and families and this aspect of the database will undoubtedly continue to provide information on the influence of ISs on genome structure, their distribution between genera and species and their degree of spread within and between ecological niches.

ACKNOWLEDGEMENTS

This site was designed by David Villa (IBCG, Toulouse) and is administered, maintained and upgraded by J.P. (IBCG) and L.L. (IBCG). It was initiated by Alain Gaekle (UCL) and Frédéric Rodriguez (IBCG) and further developed by J.P. and Philippe Azema (IBCG) with help from Michele Boschet (LMGM). We are also grateful to Robert de Boy (TIGR) for his constant supply of IS related data. The IS database was initiated by Jacques Mahillon with assistance from Réné Rezsöhazt and Bernard Hallet (UCL, Louvain-la-Neuve). It is curated by P.S. (LMGM, Toulouse), M.C. (LMGM) and J.M. with help from Jonathan File (LMGM) and Daniel De Palmenaer (UCL). Further information concerning technical aspects of this site can be obtained from P.S., J.P., M.C. or J.M. ISfinder is supported by the CNRS (France) and has received some support from l'Association de Recherche contre le Cancer (ARC, France). Funding to pay the Open Access publication charges for this article was provided by the CNRS (France).

Conflict of interest statement. None declared.

REFERENCES

1. Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A. *Mobile DNA II*. American Society for Microbiology, Washington D.C.
2. Buisine,N., Tang,C.M. and Chalmers,R. (2002) Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic *Neisseria* sp. *FEBS Lett*., **522**, 52–58.
3. De Palmenaer,D., Vermeiren,C. and Mahillon,J. (2004) IS*231*-MIC*231* elements from *Bacillus cereus sensu lato* are modular. *Mol. Microbiol*., **53**, 457–467.
4. Chandler,M. and Mahillon,J. (2002) Insertion Sequences revisited. In Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds), *Mobile DNA II*. American Society for Microbiology, Washington D.C. pp. 305–366.
5. Mahillon,J. and Chandler,M. (1998) Insertion Sequences. *Microbiol. Mol. Biol. Rev*., **62**, 725–774.
6. Campbell,A., Berg,D.E., Botstein,D., Lederberg,E.M., Novick,R.P., Starlinger,P. and Szybalski,W. (1979) Nomenclature of transposable elements in prokaryotes. *Gene*, **5**, 197–206.
7. Leplae,R., Hebrant,A., Wodak,S.J. and Toussaint,A. (2004) ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res*., **32**, D45–D49.